Entrepreneurial Experiments in Science Policy: Analyzing the Human Genome Project

Kenneth G. Huang
Assistant Professor of Management
Singapore Management University
Lee Kong Chian School of Business
50 Stamford Road #05-01
Singapore, 178899
+(65) 6828-0525
kennethhuang@smu.edu.sg

Fiona E. Murray
Associate Professor of Management
Massachusetts Institute of Technology
Sloan School of Management
50 Memorial Drive E52-567
Cambridge, MA 02142
(617) 258-0628
fmurray@mit.edu

December 2009

Forthcoming Research Policy

Entrepreneurial Experiments in Science Policy: Analyzing the Human Genome Project

Kenneth G. Huang
Singapore Management University
Lee Kong Chian School of Business

Fiona E. Murray
Massachusetts Institute of Technology
Sloan School of Management

Abstract

We re-conceptualize the role of science policy makers, envisioning and illustrating their move from being simple investors in scientific projects to entrepreneurs who create the conditions for entrepreneurial experiments and initiate them. We argue that reframing science policy around the notion of conducting entrepreneurial experiments – experiments that increase the diversity of technical, organizational and institutional arrangements in which scientific research is conducted – can provide policy makers with a wider repertoire of effective interventions. To illustrate the power of this approach, we analyze the Human Genome Project (HGP) as a set of successful, entrepreneurial experiments in organizational and institutional innovation. While not designed as such, the HGP was an experiment in funding a science project across a variety of organizational settings, including seven public and one private (Celera) research centers. We assess the major characteristics and differences between these organizational choices, using a mix of qualitative and econometric analyses to examine their impact on scientific progress. The planning and direction of the Human Genome Project show that policy makers can use the levers of entrepreneurial experimentation to transform scientific progress, much as entrepreneurs have transformed economic progress.

Keywords: Entrepreneurial Experiments, Science Policy, Human Genome Project

1. Introduction

The traditional role of science policy has been to establish and allocate government funding of scientific research. Policy makers within the key funding agencies serve as investors in the scientific community. Rather than simply responding to the supply of scientific projects, they use a variety of programmatic structures and research themes to shape both the level and direction of scientific progress. This role is justified by the long-held notion that public R&D spending should emphasize support of research in areas that are critically underinvested because they are subject to market failures (Bush, 1945; Arrow, 1962). While funding remains key to high levels of scientific output, science policy has recently been subjected to a variety of criticisms: observers have argued that the funding agencies are too conservative in their investment approach, focusing on a limited number of low-risk research projects (Kolata, 2009; Groopman, 2001). Others have pointed to the funding preferences towards older scientists with proven record of productivity thus reducing diversity (Stephan, 2008). Finally, there is limited attention paid to the diversity of the particular organizational and institutional arrangements within which scientific research is undertaken (Murray and Stern, 2007; Jones et al., 2008; Huang, 2009). Together, these criticisms point to the limited diversity of scientific research. This finding underscores the need for science policy makers and scholars to respond to recent economic theory that argues for more significant diversity in early stage research (and researchers) to ensure that the full landscape of scientific paths is explored and that suggests the importance of particular institutional choices in enabling such diversity (Aghion et al., 2008; Acemoglu, 2009; Acemoglu et al., 2009; Murray et al., 2009).

To meet the goal of increasing the diversity of scientific research, researchers and organizational arrangements, we argue that the government should re-conceptualize its role in science policy from investor to entrepreneur. Specifically, we suggest that science policy be reframed so that its core mission is to seed and support entrepreneurial experiments, encouraging the use of diverse technical, individual, organizational and institutional approaches to solve a particular problem. The experimentation perspective on entrepreneurship highlights the power of entrepreneurs to initiate a wide variety of economic experiments in the economy in order to rapidly learn about the effectiveness of different technologies, market needs and organizational arrangements (Rosenberg, 1992; Stern, 2005). While the government may not undertake all such

experiments directly, within the realm of science policy there is strong potential to act as an entrepreneur by seeding experiments and focusing proactively on assessing their results from this perspective (Greenstein 2007). Doing so, we argue, would move the government from its more typical role as a reactive investor to an entrepreneur that initiates a wide repertoire of effective interventions into the scientific community. With a proactive agenda of learning from the richly diverse set of entrepreneurial experiments, the government would also be able to promote the broader "science of science and innovation" or "science of science management" agenda. Implementing this broader agenda requires an understanding of the determinants of scientific progress and a more analytic approach to assessing the impact of technical, individual and organizational choices on scientific productivity (Lane, 2009).

In this paper, we illustrate the power of the experimentation approach to shed light on the impact of organizational diversity on scientific progress, using a large-scale entrepreneurial experiment organized by the U.S. government. While recognizing the benefits of science policy experiments ever since the Manhattan project developed the atomic bomb during World War II (Nelson, 1961), the funding orientation of the U.S. government has not been explicitly characterized as government engagement in valuable entrepreneurial experiments that the market alone would not provide nor has it been analyzed as such. In particular, by viewing each of the parallel scientific paths sponsored by government agencies (including those under the auspices of the Small Business Innovation Research (SBIR)) as an "experiment" provides a framework for analyzing how a particular scientific challenge can be more or less effectively accomplished using a variety of different technical and organizational choices. This in turn deepens our understanding of the link between organizational arrangements and scientific productivity.

The entrepreneurial experiment we explore in this paper is the Human Genome Project (HGP), (or more precisely the Human Genome *Projects*) funded by the United States Department of Energy (DOE)[1] and the National Institutes of Health (NIH)[2] , as well as the Wellcome Trust in

[1] After the atomic bomb was developed and used, the U.S. Congress charged DOE's predecessor agencies (the Atomic Energy Commission and the Energy Research and Development Administration) with studying and analyzing genome structure, replication, damage, and repair and the consequences of genetic mutations, especially those caused by radiation and chemical by-products of energy production. From these studies grew the recognition that the best way to study these effects was to analyze the entire human genome to obtain a reference sequence. Planning began in 1986 for DOE's Human Genome Program and in 1987 for the National Institutes of Health's

the United Kingdom. While typically regarded as one monolithic science project, in fact this massive effort to sequence the entire human genome was carried out in seven public research centers each with different organizational arrangements. Moreover, about eight years after the public Projects' initiation, start-up Celera Genomics began a separate, privately funded quest to complete a full genome sequence, using an alternative technical approach and carried out with an entirely distinctive organizational model: both the organization of the work and the institutions governing data access contrasted sharply with the public Projects.

The remainder of this paper proceeds as follows: In Section 2 we provide a deeper understanding of the nature of entrepreneurial experiments and their application to science policy. In Section 3 we then use this framework to describe the Human Genome Project(s) as an entrepreneurial experiment. In Section 4 we analyze the impact of different organizational choices on the productivity of the different HGP groups illustrating the potential for program evaluation of different experiments. In Section 5 we provide a broader framework for the design and evaluation of economic experiments in the science policy setting.

2. Economic experimentation

We are all familiar with the central role of scientific experimentation in the pursuit of technical progress; it has become a foundational tenet of progress (Merton 1966) not least because even with the most detailed theoretical models, it is rarely possible to predict *ex ante* the most appropriate research line an advance of an experiment. While scientific or technical experiments are widely understood, economic experiments are harder to envision. An economic experiment can be defined as the choice of a particular combination of technical, market and economic characteristics that form the basis of an opportunity that will hopefully create value and

program. The DOE-NIH U.S. Human Genome Project formally began October 1, 1990, after the first joint 5-year plan was written and a memorandum of understanding was signed between the two organizations.

[2] The National Institutes of Health (NIH), founded in 1887, is one of the world's premier medical research centers, and the federal focal point for medical research in the U.S. The NIH, comprising 27 separate Institutes and Centers, is one of eight health agencies of the Public Health Service which, in turn, is part of the U.S. Department of Health and Human Services. The primary mission of NIH is to "acquire new knowledge to help prevent, detect, diagnose, and treat disease and disability, from the rarest genetic disorder to the common cold…[and] to uncover new knowledge that will lead to better health for everyone." By its key involvement in the HGP, NIH works toward that mission and advances human health by "conducting research in its own laboratories; supporting the research of non-Federal scientists in universities, medical schools, hospitals, and research institutions throughout the country and abroad; helping in the training of research investigators; and fostering communication of medical and health sciences information."

economic gain (Rosenberg 1992). With our focus on experiments designed to increase the degree of scientific productivity (rather than on economic value per se), we use the term entrepreneurial experiment because as Stern (2005, p. 16) notes, "*While economic experiments can be (and are) implemented in established companies (and can even be found in the public sector), economic experimentation is at the heart of the entrepreneurial process.*" Thus we can consider experiments in science policy as key entrepreneurial experiments.

In the realm of science policymaking and the allocation of government research funding, we argue for the critical importance of entrepreneurial experiments expanding, varying and testing the causal impact of different technical, organizational and institutional arrangements on the creation of scientific value. This follows from the view that experimentation should focus not only on generating information about the best technical path but also determine the best organizational or institutional approach – in much the same way that companies experiment with the most effective market application or business configuration (Greenstein, 2007). The analogy is simple: scientists might use economic experiments to reduce the uncertainty about the way in which particular factors increase or decrease their probability of success. These factors can involve particular combinations of technical approaches, but they can and should also be organizational. Although some argue that science cannot be "managed" and is a black box inside which "unmanageable" individuals ply their craft, evidence suggests that specific interventions in organization, incentives, governance do in fact shape scientific productivity as do broader institutional interventions such as ownership, sharing and exchange (Furman and Stern 2006; Henderson and Cockburn, 1994; Murray and Stern, 2007; Huang and Murray, in press; Huang, 2009). If these interventions do in fact shape the outcome of scientific projects, then opportunities for economic experiments abound well beyond the traditional technical domain. The government is well placed to serve as an entrepreneur in seeding and promoting these experiments, thus increasing the diversity of scientific research along many dimensions.

Entrepreneurial experiments are of potentially significant value because, as Rosenberg (1992, p. 181) has persuasively argued, "*The freedom to conduct experiments is essential to any society that has a serious commitment to technological innovation or to improved productive efficiency….Only the opportunity to try out alternatives, with respect both to technology and to*

form and size of organization, can produce socially useful answers to a bewildering array of questions that are continually occurring in industrial (and in industrializing) societies." By creating the conditions for economic experimentation (by entrepreneurs and others), this approach also has the potential to drive much greater technical, market and organizational diversity into the innovation system. For example, early entrepreneurs providing Internet services engaged in a variety of market experiments and in organizational experiments around how to construct the value chain for effective competitive advantage (Greenstein, 2007), and similar patterns of experimentation appeared among early dot-com start-ups (Goldfarb et al., 2007). Even among those attempting to monetize and seek economic value based on the early developments in biotechnology we see a wide range of market, organizational and institutional experiments (Kaplan and Murray, in press).

While entrepreneurs have high incentives to engage in economic experiments, recent economic theory has argued that current incentive systems push scientists to follow a too narrow set of potential paths or research lines (Acemoglu, 2009), leading to insufficient diversity in the scope of early stage R&D projects. This gap in experimentation suggests a potentially important role for science policy makers within the government (as well as not-for-profits) to serve as entrepreneurs, creating the conditions for economic experiments linked not simply to immediate value-creating outcomes but also to the productivity of scientific knowledge. Such diversity and experimentation must highlight not only the technical dimension experimentation but also along the market and organizational dimensions. Such an approach is particularly salient in R&D because, as a highly risky endeavor subject to uncertainty in outcomes (Nelson, 1959), it requires government effort to ensure the appropriate level investment in R&D. It also suggests that government explicitly view themselves as providing diverse types of investment that the market would not otherwise provide (Nelson, 1961).

Government efforts at parallel R&D funding are not entirely new. As Nelson (1961, p.353) quoted in his paper on parallel research, when James Conant wrote to Vanevaar Bush in 1942 regarding the parallel experimental approaches to the Manhattan Project *"All five methods will be entering very expensive pilot plant development during the next six months. . . . [But] while all five methods now appear to be about equally promising, clearly the time to production . . . by*

the five routes will certainly not be the same but might vary by six months or a year because of un- foreseen delays. Therefore, if one discards one or two or three of these methods now, one may be betting on the slower horse unconsciously."

More contemporary examples of current government efforts in furthering experimentation include the SBIR Program (see Link and Scott in this volume for a thorough analysis of experiments in this program) and a variety of competition-based procurement processes such as the Advanced Technology Program (Link and Scott, 2001) now transformed into the Technology Innovation Program. For example, the 2009 Technology Innovation Program (TIP) competition held by the U.S. National Institute of Standards and Technology (NIST) highlighted, among other areas, a competition for proposals "accelerating the incorporation of materials advances into manufacturing processes" (National Institute of Standards and Technology, 2009). This requires a variety of organizational approaches including diverse collaborations and partnerships. Similarly, the recent the surge of interest in R&D prizes by the government also provides another potential mechanism for diversity and experimentation along a variety of dimensions (Horrobin, 1986; Kalil, 2006).

While these specific examples hint at the government's potentially powerful role in shaping entrepreneurial experiments, a framework for the consistent and thorough analysis of these experiments is missing. In the realm of R&D and the funding of scientific research projects, we argue that the government has an important opportunity to structure its experimentation along three key dimensions: technical, individual and organizational. By widening the scope of approaches to a particular research challenge, critical diversity can be introduced into the system and experimentation can entail exploring multiple lines research (Murray et al., 2009). Such efforts can potentially overcome strong criticism directed towards government funding agencies for their low-risk, conservative approach towards funding of research in areas such as cancer (Kolata, 2009; Groopman, 2001). In addition, by broadening the range of individuals and organizations participating in R&D and receiving funding, such as through new investigator programs or collaborative approaches, individual diversity can also contribute to experimentation (Jones et al., 2008). Lastly, as Foray (2000, p.1) has noted, "policy makers must themselves be willing to experiment with new institutional arrangements," reminding us that entrepreneurial

experiments undertaken in the policy domain should attend to the diversity of organizational forms (Stern, 2005). By more explicitly experimenting with the organization of research, the government can add variety of a different sort, one that reflects the growing variety of organizational approaches to the kind of creative work that is coming to dominate our economy (Bechky and Hargadon, 2007). Specifically, while some evidence exists that particular organizational choices are correlated with more productive scientific outcomes, the current range of organizational experiments is limited. Likewise, with regards to institutional arrangements, heated debate has arisen over the role of IP rights governing particular aspects of scientific knowledge (Heller and Eisenberg, 1998; Walsh et al., 2005; Murray and Stern, 2007; Huang and Murray, in press) suggesting that institutional experimentation must continue. However, in the absence of more carefully designed entrepreneurial experiments followed by careful analysis, we cannot have definitive answers that allow for thorough policy guidance. We illustrate the value of entrepreneurial experimentation in the area of scientific research by examining the sequencing of the entire human genome in the context of distinctive organizational and institutional choices. Although these choices are not as fully documented as one might prefer in a consciously designed experiment (and have not been assessed using recent advances in program evaluation), we capture and analyze their impact on a variety of outcome measures. These measures include commercialization activities, technology transfer decisions and broad impact, as well as measures of knowledge production, its diffusion, accumulation and translation into commercial outcomes.

3. The Human Genome Project as an entrepreneurial experiment

3.1. Organization of the Human Genome Project and historical overview

The Human Genome Project was a 13-year, $3.8 billion research effort funded and coordinated by the U.S. Department of Energy and the National Institutes of Health. It was also the most expensive and arguably the most significant life science research project undertaken in the history of U.S science. The human genome is the entirety of hereditary information of Homo sapiens, stored in 23 chromosome pairs and contains approximately 23,688 protein-coding genes. The explicit goals of the HGP were to identify all 23,688 genes in the human DNA, sequence its 3 billion nucleotide base pairs (adenine, guanine, cytosine, thymine abbreviated as A, G, C, T

respectively), store this information, develop the methods and tools for data analyses and transfer related technologies to the private sector. The HGP grew from an ambitious idea in the minds of scientists, legislators and government agencies to a set of landmark scientific achievements that redefined the life sciences landscape, gaining broader public support along the way. It culminated in the successful sequencing and publication of the draft sequence of the human genome in 2001 and the complete sequence in 2003.[3]

As an entrepreneurial experiment organized by the government, the different stakeholders that comprised the HGP exemplified the diversity in talent, approach and organization. The seven major public genome centers, led by the U.S.-based Whitehead Institute for Genome Research at Massachusetts Institute of Technology in Cambridge Massachusetts, and the Wellcome Trust Sanger Institute in Cambridgeshire, U.K., also included Washington University Genome Sequencing Center in St. Louis Missouri, Baylor Human Genome Sequencing Center in Houston Texas, University of Washington Genome Center in Seattle Washington, Stanford Human Genome Center in Palo Alto California, and the Department of Energy Joint Genome Institute. These centers made up the main thrust of the public effort by completing altogether more than 99% of the sequencing. The private effort was spearheaded by Celera Genomics, which started competing against the public project in 1998. While these centers all engaged in aspects of the same overall project, they differed in their origins, organizational characteristics, size, talent pool, and incentives, particularly choices governing disclosure and intellectual property (IP) policies which varied sharply between the public and private efforts but even within the public project depending on the local institutional rules surrounding intellectual property rights.

Before examining the specific variations along different dimensions of each genome center and the competition between the public vs. private sequencing efforts, it is important to first understand the major historical and scientific milestones of the HGP and the role of government in organizing and funding the genome centers to ensure early completion of this mammoth project.

[3] DOE Major Timeline: http://www.ornl.gov/sci/techresources/Human_Genome/project/timeline.shtml

The first serious push to sequence the human genome began in 1984, when Robert Sinsheimer, a distinguished molecular biologist and senior administrator at the University of California, proposed to the University of California President David Gardner that an institute be established for this purpose on the University of California Santa Cruz (UCSC) campus. Although the proposal was not funded, Sinsheimer continued the discussion with other molecular biologists at UCSC then including Harry Noller, Robert Edgar, and Robert Ludwig.

Sinsheimer held a meeting of researchers in UCSC in May 1985, proposing that the entire human genome should not only be mapped with scattered but specific "road-markers", but also sequenced to determine the order of each A, G, C and T. A number of distinguished biologists active in genetics and gene mapping were present during the meeting, including Harry Noller, now the Sinsheimer Professor of Molecular Biology at UCSC; geneticist David Botstein, now the director of the Lewis-Sigler Institute for Integrative Genomics at Princeton University; Leroy Hood, now the President of the Institute for Systems Biology in Seattle; and the 1980 Nobel laureate in Chemistry, Walter Gilbert of Harvard University, who later became an enthusiastic advocate of this notion. At around the same period of time, other scientists made independent proposals for the sequencing of the human genome, notably Renato Dulbecco of the Salk Institute and Charles DeLisi of the U.S. DOE.

The proposal seemed idealistic and almost logistically impossible (Lander and Weinberg, 2000) – the human genome encompasses about 3 billion bases of DNA and the technology then only allowed reading lengths of about 300 bases in each analysis. Decades of work by a huge number of research scientists and technicians would be required to perform and complete the job. In addition, analysis of a single base would cost over $10 at that time and it often required more than a day to sequence 50 to 100 bases. Such suggestion to sequence the genome then was not only ambitious but also prohibitively expensive.

Furthermore, opponents argued that sequencing the human genome would be a vast waste of effort as majority of it, maybe as high as 95%, does not encode useful protein or regulatory information, known as "junk DNA". Such enormous effort to obtain detailed sequence

information about DNA, as they argue, would have little hope of shedding useful insight into biological function. However, the proposal prevailed.

From early 1986, the government officially started a series of important initiatives and meetings to organize this major effort. From the Office of Health and Environmental Research (OHER) of U.S. DOE,[4] (now the Office of Biological and Environmental Research (BER)), biophysicist and administrator Charles DeLisi[5] joined David A. Smith[6], director of the DOE Human Genome Program, to organize and hold a conference in Santa Fe. The goal of the meeting, a follow-up to the previous Santa Cruz meeting, was to assess the feasibility of a human genome initiative and the role of DOE in sequencing the entire human genome,. Following the Santa Fe conference, OHER of DOE, announced the Human Genome Initiative. With $5.3 million, pilot projects first began at DOE national laboratories to develop critical technologies and resources. Three genome research centers were established between 1988 and 1989 at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL).

Legislation to begin planning a mammoth project, resulting in a detailed analysis of all the genes in the human body, was introduced in the Senate in December 1987, sponsored by Senator Pete Domenici (R-N.M.), Edward M. Kennedy (D-Mass.), Lawton Chiles (D-Fla.), Patrick J. Leahy (D-Vt.) and Bob Graham (D-Fla.). The bill focused on establishing a freestanding National Biotechnology Policy Board and Advisory Panel in order to ensure a high level of competitiveness for the biotechnology industry in the U.S.; these groups would foster policies to "enhance the efficient and timely advance of basic and applied biotechnology-related research".

---

[4] After the atomic bomb was developed and used, the U.S. Congress charged DOE's predecessor agencies (the Atomic Energy Commission and the Energy Research and Development Administration) with studying and analyzing genome structure, replication, damage, and repair and the consequences of genetic mutations, especially those caused by radiation and chemical by-products of energy production. From these studies grew the recognition that the best way to study these effects was to analyze the entire human genome to obtain a reference sequence. Planning began in 1986 for DOE's Human Genome Program and in 1987 for the National Institutes of Health's program. The DOE-NIH U.S. Human Genome Project formally began October 1, 1990, after the first joint 5-year plan was written and a memorandum of understanding was signed between the two organizations.

[5] DeLisi, Charles (July 2001). "Genomes: 15 Years Later. A Perspective by Charles DeLisi, HGP Pioneer." *Human Genome News*, Vol.11, No. 3-4.
http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v11n3/05delisi.shtml

[6] Smith, David A. (September-December 1995). "Evolution of a Vision: Genome Project Origins, Present and Future Challenges, and Far-Reaching Benefits". *Human Genome News*, 7(3-4): 2.
http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n3/02smithr.shtml

The bill also centered on a huge project to map and sequence the human genome. However, the bill only authorized enough funds to support the Board through 1993 (The Washington Post, 1987).

In the same year, congressionally chartered DOE advisory committee, Health and Environmental Research Advisory Committee (HERAC),[7] boldly recommended a 15-year, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE also designated important multidisciplinary human genome centers in the U.S. that year, including the *Whitehead Institute for Genome Research at Massachusetts Institute of Technology, Washington University School of Medicine Genome Sequencing Center*, and *University of Washington Genome Center*. National Institute of General Medical Sciences (NIGMS) at National Institutes of Health started funding the genome projects that year.

In 1988, congressional Office of Technology Assessment (OTA) and National Academy of Sciences (NAS) National Research Council (NRC) committees recommended a concerted genome research program. Over the next year, DOE and NIH held several other meetings and independent hearings by OTA and by NAS to discuss the plans for the HGP. Also in the same year, the Human Genome Organization (HUGO)[8] was founded by scientists to coordinate efforts internationally. The first annual Cold Spring Harbor Laboratory meeting on human genome mapping and sequencing was also held. That year, Congress funded both the DOE and the NIH to start further exploration of the human genome, and DOE and NIH signed a formal Memorandum of Understanding (U.S. Department of Energy, 1990a), which outlined plans for cooperation on genome research to "coordinate research and technical activities related to the human genome".

These activities and the reports published culminated in 1988 when the government established the Genome Office at the National Institutes of Health and appointed Nobelist James Watson as its head (The Washington Post, 1989). This subsequently became the National Center for Human Genome Research (NCHGR) in October 1989 by Congressional authorization to carry out the

[7] http://www.ornl.gov/sci/techresources/Human_Genome/project/herac2.shtml
[8] http://www.gene.ucl.ac.uk/hugo/

role of NIH in the HGP. NCHGR was created to support the development of resources and technology that would accelerate genome research and its application to human health. In 1997, the NCHGR at NIH was restructured and renamed the National Human Genome Research Institute (NHGRI) by the U.S. Department of Health and Human Services (DHHS) to allow NHGRI to operate under the same legislative authorities as other NIH research institutes. It became one of the 27 institutes and centers that make up of NIH. Advisory boards were created to serve both the NIH and DOE.[9] Meetings were held twice annually on succeeding days since several committees were "joint", especially the one on data (Joint Informatics Task Force).

In 1989, the biology community called for a $3 billion project to identify and decipher each of the (then) estimated 30,000 genes that were understood to govern the form and function of the human body (Chicago Sunday Times, 1989).[10] The largest funding agency of such activities, NHGRI, was funded yearly through Congressional appropriation through a standard budget process.[11] In turn, the NHGRI was guided by a series of five-year plans outlining the priorities and goals of the project. These plans detailed the objectives of the program to the scientific community and informed the public while ensuring measurable aims to steer the work and determine NHGRI's progress. In allocating funds, NHGRI published its areas of research interest in program announcements so that individual scientists or academic institutions, non-profit organizations, community hospitals and companies could apply for research funding. A two-tier, peer-review process evaluated all applications and NHGRI funded the highest ranked

[9] http://www.genome.gov/10000905

[10] The Human Genome Project is sometimes reported to have a cost of about $3 billion. However, this figure refers to the total projected funding over a 13-year period (1990–2003) for a wide range of scientific activities related to genomics. These include studies of human diseases, experimental organisms (such as bacteria, yeast, worms, flies, and mice); development of new technologies for biological and medical research; computational methods to analyze genomes; and ethical, legal, and social issues related to genetics. Human genome sequencing represents only a small fraction of the overall 13-year budget.

[11] Every year, the President submits a budget request for the entire federal government to Congress, which then conducts hearings on that budget request. Different committees have the authority to approve specific sections of the federal budget. Representatives of NHGRI testify before the House and Senate subcommittees on Labor; Health and Human Services; Education; and related agencies, where Congress is updated on the accomplishments, needs and opportunities of NHGRI. Congress also hears testimony from public witnesses such as experts in genetic research, or representatives of genetic disease advocacy groups.

After listening to the testimony, the House of Representatives determines a funding level for NHGRI and sends its recommendation to the Senate. After the Senate conducts hearings, both bodies of Congress meet to agree on funding levels for all of the institutes and centers of the National Institutes of Health (NIH), including NHGRI. Congress then sends an appropriation bill with the recommended funding levels to the President. After the President signs the budget, NHGRI receives its funding.

See http://www.house.gov/rules/budget_pro.htm and http://www.genome.gov/10000933

proposals that were within the program priorities.[12] NHGRI eventually funded the various genome centers in the U.S..[13]

These events culminated in the 5-year U.S. Human Genome Project (HGP) plan jointly presented by DOE and NIH to the Congress in 1990. The report scrutinized the present state of genome science and detailed the complementary approaches of the two agencies for attaining scientific goals, while laying out concrete plans for governing research agendas. The report also explained the collaborative effort among U.S. and international agencies and presented the budget of "…about $200 million per year for approximately 15 years" (U.S. Department of Energy, 1990b). The estimated 15-year project formally began in that year. Several projects had already begun to mark gene sites on chromosome maps as sites of mRNA expression, while research and development were also underway for efficient production of more stable, large-insert Bacterial Artificial Chromosomes (BACs), which were essential tools for constructing genomic maps.[14] The promise and benefits of the now formalized Human Genome Project started to receive wider media attention in major newspapers and magazines such as the Washington Post (1990), the Wall Street Journal (1990) and Business Week (1990).

In 1991, the human chromosome mapping data repository, Genome Database (GDB),[15] an international collaboration in support of the human genome project, was established. RTI International, a prominent research institute located in the Research Triangle Park in North Carolina, hosted it. During the same period, the technology and science were improving dramatically to produce high cost reduction in sequencing efforts. For example, the cost of sequencing 10,000 bases in a single day has dropped to about a dollar a base (Hunkapiller et al, 1991). By 1993, many laboratories concurrently used robotic analyzers to sequence 500,000 bases daily and they cost about $0.10 to $0.15 a base.

---

[12] NHGRI Budget and Financial Information, 2004. http://www.genome.gov/10000933

[13] For example, the Whitehead Institute for Biomedical Medical Research, Cambridge, Massachusetts, received approximately $35 million from the NHGRI of NIH, to participate in the first year of the full-scale effort to sequence the human genome. NHGRI also funded Washington University Genome Sequencing Center, Baylor College of Medicine Human Genome Sequencing Center, University of Washington Genome Center, and the Stanford Human Genome Center. In contrast, the Sanger Institute in the U.K. is funded primarily by the Wellcome Trust.

[14] Bacterial Artificial Chromosomes (BACs) comprised one of the most utilized resources .

[15] http://www.gdb.org/

The ensuing years saw continued and promising development and improvement of resources, technologies and scientific techniques, and optimism for the potential for commercialization of the human genome sequences (Business Week, 1992). In 1992, low-resolution genetic linkage map of entire human genome was published. The guidelines for data release and resource sharing of the human genome project were also announced by DOE and NIH to encourage data and resource sharing (U.S. Department of Energy, 1993).

In 1993, the international Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium was established to coordinate efficient mapping and sequencing of gene-representing cDNAs (U.S. Department of Energy, 1995a). The consortium produced highly cited scientific papers that made significant contribution to the progress of the Human Genome Project (The Scientist, 1999). At the same time, DOE and NIH revised their initial 5-year plan for the Human Genome Project due to rapid advances in genome research and more in-depth understanding of how to attain long-term objectives (Collins and Galas, 1993). In terms of technology advances, Lawrence Berkeley National Laboratory (LBNL) of DOE implemented a novel transposon-mediated chromosome-sequencing system while Gene Recognition and Analysis Internet Link (GRAIL) sequence-interpretation service maintained by Oak Ridge National Laboratory (ORNL) of DOE started to provide Internet access. [16]

1994 brought in encouraging news that the genetic-mapping 5-year goal presented by DOE and NIH was achieved one year ahead of schedule (U.S. Department of Energy, 1994). The second-generation DNA clone libraries representing each human chromosome were also completed that same year by Lawrence Livermore National Laboratory (LLNL) and Lawrence Berkeley National Laboratory (LBNL).

In terms of scientific breakthroughs, LANL and LLNL respectively announced in 1995 the completion of high-resolution physical maps of chromosome 16 and 19 (U.S. Department of Energy, 1995b). Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps were also published (U.S. Department of Energy, 1995c). Research led by scientists from the *MIT Whitehead Institute Center for Genome Research* and Genethon revealed and published the

[16] http://genome.ornl.gov/

physical map of the human genome with more than 15,000 sequence tagged site (STS) markers (U.S. Department of Energy, 1996a).

In 1996, DOE and National Center for Human Genome Research (NCHGR) at National Institutes of Health issued human subject guidelines for large-scale sequencing projects (U.S. Department of Energy, 1996b). Another landmark event that occurred during that year is the large-scale sequencing strategy meeting for international coordination of human genome sequencing in Bermuda held from 25 to 28 February, 1996. It was sponsored by the *Wellcome Trust, U.K. Medical Research Council*. About 50 scientists from countries publicly supporting large-scale human genome sequencing attended the conference. The conference was designed to coordinate, compare, and evaluate human genome mapping and sequencing strategies; consider the potential role of new technologies in sequencing and informatics; and discuss scenarios for data release. A consensus was reached that the eventual sequencing outcome representing the first human genome sequence should be conducted at a high degree of accuracy.

Participants (participating organizations and funding agencies) in the HGP agreed on sequencing data release policy at the Second International Strategy Meeting on Human Genome Sequencing in 1997.[17] That year, in order to implement high-throughput activities, DOE also formed the *Joint Genome Institute (JGI), an* effort to tie the expertise and resources in genome mapping, DNA sequencing, technology development, and information sciences pioneered at DOE genome centers: Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL) by forming*)*. The institute would start with sequencing and functional genomics (U.S. Department of Energy, 1996).

3.2. Private Celera Genomics versus public Human Genome Sequencing Consortium

June 1998 marked the moment when a true market-driven entrepreneurial experiment joined the previously government dominated efforts in human genome sequencing; a competing private quest to sequence the human genome was launched by *Celera Genomics* (co-founded by Applera

[17] Held in Bermuda from 27th February 1997 to 2nd March 1997.

Corporation (then called PE Corporation) and Dr. J. Craig Venter).[18] The effort was also headed by Venter[19] who had been a scientist at the NIH during the early 1990s, when the HGP was started, . He left NIH to start the Institute for Genomic Research (TIGR) in 1992, which led the first successful sequencing of an entire organism's genome (*Haemophilus influenzae* bacterium). Celera's mission was to generate and commercialize genomic information. Specifically, it aimed to sequence the entire human genome and provide its putative clients, pharmaceutical and biotechnology companies, with early access to the resulting data. The firm aimed to complete the sequencing of the human genome at a faster pace (within three years) and at a fraction of the cost of the publicly funded project ($300 million vs. $3 billion). Figure 1 illustrates the organization of the HGP, mapping the relationship among the key stakeholders and competitors.

Insert Figure 1 about here

When Celera entered the competition, more than eight years into the Human Genome Project, it had the advantage of freely obtaining the existing publicly available data from GenBank (nearly one-third of the human genome sequence in finished or draft form), as well as building on the technical groundwork laid by the public consortium. Using state-of-the-art sequencing technology supplied by Applied Biosystems Group of the Applera Corporation and sophisticated internally-developed informatics, it pioneered a technique, whole genome shotgun sequencing, that had been used to sequence bacterial genomes of only up to six million base pairs in length, far fewer than the three billion base pair human genome.

Collaboration and pooling of efforts with the public consortium would facilitate faster completion of the sequencing project, and win Dr. Venter more friends in the academic scientific community from which he came. [20] However, this option disappeared, pardly due to rival agendas; in particular, there was disagreement about access to sequencing data (Venter refused to

[18] Celera Genomics belonged to the Applied Biosystems Group business unit of the Applera Corporation until it was spun off in July 2008 to become an independent publicly traded company.
[19] Dr. Craig Venter headed Celera from its founding to early 2002, when he was fired due to a conflict with the main investor, Tony White who had also been with the company since its founding.
[20] In fact, according to an interview on November 14, 1999, Dr. Paul Gilman, a senior executive at Celera then, commented that Dr. Venter was open to the idea of collaboration with the public consortium but that no specific proposal was under discussion.

deposit Celera's data in the unrestricted public database GenBank) and about the direction of the project.

In terms of disclosure and IP, Celera initially announced that it would seek patent protection on "only 200 to 300" genes, but later amended this to seek "intellectual property protection" on "fully-characterized important structures" (or any commercially valuable DNA sequences) amounting to 100 to 300 targets. The firm eventually filed preliminary ("place-holder") patent applications on 6,500 whole or partial genes (BBC News, 1999). In fact, since Celera's founding in 1998 to the completion of the HGP in 2003, it had filed 259 granted genomics, methods and tools patents according to the United States Patent and Trademark Office (USPTO). Out of which, at least 65 patents claimed part of a gene or gene sequences (Jensen and Murray, 2005). While the company's initial policy was to hold back data from sections of the genome they sequenced for commercial exploitation, Celera later amended its policy in response to critics and promised to publish their findings in accordance with the terms of the 1996 "Bermuda Statement",[21] by releasing new data annually (while the publicly funded HGP released its new data daily). Nevertheless, it would not permit free redistribution or commercial use of the data and had set a maximum threshold for amount of sequence data a researcher could download at any given time, unlike the publicly funded project. Celera also planned to profit from its sequencing effort by establishing a value-added database of genomic data that users could subscribe to for a fee. Essentially, Celera had incorporated the public data into their genome while restricting public use of Celera data.

In March 2000, President Clinton announced that the genome sequence could not be patented, and should be made freely available to all researchers (although it was not clear if it was binding). The statement sent Celera's stock plummeting and dragged down the biotechnology-heavy Nasdaq. The biotechnology sector lost about $50 billion in market capitalization within two days.

The "working draft" DNA sequence of the human genome was jointly announced in June 2000 by the HGP leaders Ari Patrinos (director of DOE Human Genome Program and Biological and Environmental Research Program) and Francis Collins (director, NIH National Human Genome

[21] http://en.wikipedia.org/wiki/Bermuda_Principles

Research Institute), as well as Craig Venter (head of Celera Genomics) and then U.S. president Bill Clinton. The International Human Genome Sequencing Consortium, led by NHGRI and DOE and Celera, soon followed by publishing details of their drafts in February *2001*. A special issue of *Nature* (International Human Genome Sequencing Consortium, 2001) on 15 February 2001 published the public consortium's scientific results one day ahead of Celera's publication in *Science* (Venter et al., 2001). These papers described the methods used to produce the draft sequence and offered analysis of the sequence. These drafts covered about 83% of the genome (90% of the euchromatic regions with 150,000 gaps and the order and orientation of many segments not yet established). In February 2001, at the time of the joint publications, press releases announced that the project had been completed by both groups.

The fierce private versus public competition had spurred the publicly funded genome centers to modify their strategy in order to accelerate progress. The *Wellcome Trust Sanger Institute* from the public consortium had completed the sequencing of chromosome 20 by 2001.[22] This was the third (following chromosomes 22 and 21) and the largest of the human chromosomes to be completed to the high scientific standard specified by the Human Genome Project. On April 14, 2003, the public consortium announced the successful completion of the Human Genome Project more than two years ahead of schedule (Reuters Health, 2003).[23] In the same year, the finished sequences of chromosomes 14 (January),[24] Y (June),[25] 7 (July)[26] and 6 (October)[27] were published. These were followed by the finished sequences of chromosomes 13[28] and 19[29] published in April 2004.

3.3. The genome centers and their organizational choices governing disclosure and IP

While the public genome centers within the International Human Genome Sequencing Consortium had similar goals – scientific sequencing and data release standards – they varied along specific dimensions. Table 1 summarizes the key attributes of each of the seven public

[22] *Nature*, December 20, 2001.
[23] Also see: http://www.ornl.gov/sci/techresources/Human_Genome/project/50yr/press4_2003.shtml
[24] *Nature*, January 2003
[25] *Nature* 423**,** 810-813 (19 June 2003)
[26] *Nature* 424, 157-164 (10 July 2003)
[27] *Nature* 425, 805-811 (23 October 2003)
[28] *Nature* 428, 522-528 (01 April 2004)
[29] *Nature* 428, 529-535 (01 April 2004)

genome centers in terms of founding year, leadership, affiliations, percentage of human genome sequence completed, chromosomes sequenced, operating budget, funding or employment size.

Insert Table 1 about here

We now focus on the variations in the different organizational choices governing disclosure and IP policy for each center. These variations enable a more precise analysis of the economic experiment organized by NHGRI and DOE (to be described in the Section 4).

The Whitehead Institute/ MIT Center for Genome Research (WIBR) chose to actively patent and license its genome research outputs. However, it practiced forms of data release in accordance with the data release policy endorsed by both the Wellcome Trust Sanger Institute and NIH (Bentley, 1996).[30] This formed an interesting and contrasting variation with most of the other genome centers in terms of how it governed information disclosure, production and dissemination of scientific knowledge and intellectual property rights for commercialization of biotechnologies.

In contrast, the Wellcome Trust Sanger Institute was one of the major proponents of the "no patent" policy, believing *in the immediate and free release of genomic sequence information to the public domain.* This policy (i) permits coordination; (ii) is of immediate value to others and is not misleading; and (iii) promotes maximum accessibility of the human genome sequence for interpretation and exploitation. Furthermore, the Institute believes that such "activities should flourish in both the academic and commercial sectors…." and "withholding the genomic sequence ingredient from any academic or commercial laboratory with such knowledge impedes scientific progress and is not in the international public interest." It has taken the stand that "patenting of raw human genomic DNA sequence or partial or complete gene sequences of unknown function is inappropriate," as it may "discourage further research and development by others, for fear that future inventions downstream of the gene sequence itself could not be adequately protected… Free release of sequence data will also encourage exploitation by a

[30] This paper was written on behalf of the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, and Genome Sequencing Center, Washington University, St. Louis.

maximum number of commercial and academic centers that are keen to compete in the development of new therapeutic agents." (Bentley, 1996). This serves as a sharp contrast to the patent policy adopted by the Whitehead Institute.

The Washington University Genome Sequencing Center (WUGSC), like the Sanger Institute, was a champion of the non-patenting policy on all its HGP research output, believing this choice will maximize the dissemination and utilization of knowledge and commercialization (Bentley, 1996). *On a much smaller scale than the MIT Whitehead Institute, the Baylor College of Medicine Human Genome Sequencing Center* (BCM-HGSC) *selectively patents* some of the related technologies derived from the sequencing of the human genome.

Although *University of Washington Genome Center* (UWGC) "emphasizes innovative technology development and high quality sequence production", it "*decided to not to patent any of the sequences* and it *holds no patents on anything*…"[31] This was in line with the non-patenting policy of the Sanger Institute and the Washington University Genome Sequencing Center. Like most other human genome centers, *Stanford Human Genome Center* (SHGC) follows a non-patenting policy on its HGP research output. Finally, *DOE Joint Genome Institute*, like most its counterparts in the HGP including the SHGC, adopts a rapid disclosure of information and non-patenting policy on its HGP research output.

4. Analyzing the Human Genome Project experiments

4.1. Data and Measures

Economic experimentation by government in large-scale scientific projects such as the HGP introduces diversity and variety on different dimensions (e.g. talent, approach and disclosure incentives) as described in the previous sections. These variations produced by government's experimentation, specifically on organizational disclosure and IP policy, provide an "experiment" in which to analyze the impact of a pro-patenting policy vs. full and rapid disclosure of scientific knowledge in the genome centers. Table 2 summarizes the organizational

[31] According to informal interview with the Computer Support Analyst from the University of Washington Genome Center (2003).

choices in disclosure and patenting policy of the seven public genome centers vs. private Celera Genomics. The Whitehead/MIT Center and Baylor Human Genome Sequencing Center, from the public consortium, and the corporate Celera Genomics allowed for patenting. In contrast, the other five public genome centers, namely Sanger Institute (U.K.), Washington University Genome Sequencing Center, University of Washington Genome Center, Stanford Human Genome Center and DOE Joint Genome Institute, established a full disclosure and strictly no patenting policy.

Insert Table 2 about here

These seven public centers provide a matched and well-controlled opportunity to examine the impact of patent and disclosure policy on scientific knowledge produced by the genome centers. The heterogeneity of genomic research data is reduced by focusing only on knowledge related to human genome sequencing. In addition, these centers do not have knowledge *ex ante* on the choice and fruitfulness of the pieces of chromosomes they sequence. All the public centers, as part of the same consortium, follow similar research and publication strategy (although not IP and disclosure policy) in the collaborative sequencing of the human genome.

To understand the effects of IP and disclosure policy on knowledge production and commercialization by the genome centers, we collected data on all publications, patents and commercialization efforts from the HGP by the seven public genome centers as well as private Celera Genomics for comparison. Table 3 describes the variables in this study. Table 4 provides the summary statistics while Table 5 shows the correlation matrix for the key variables.

Insert Table 3 about here

Insert Table 4 about here

Insert Table 5 about here

The **publication data** is obtained from the ISI Web of Science, which provides the most comprehensive coverage of peer-reviewed scientific research articles available. These publications are cross-checked with the publication records announced by each genome center or listed on their website. This yielded a total of 1484 publications from January 1990, the year in which the HGP started, to December 2003, the year it ended. The number of publications across the seven public genome centers and Celera Genomics shows a generally increasing trend from 1990 to 2003, all peaking at 2002 or 2003 (with the exception of Baylor which peaked at 1995), near or at the completion of the HGP, when most results are published (see Figure 2).

Insert Figure 2 about here

The ***dependent variable*** is the *number of cumulative forward citations* to each scientific publication. We use publication citations to each genomic paper (i.e. peer-reviewed publications citing the focal paper) as a proxy for the importance of the scientific knowledge in the form of follow-on knowledge accumulation. Our citation-based approach follows a long literature using citations to trace the flow of ideas and their follow-on accumulation in later knowledge production (de Solla Price, 1965; Hall, Jaffe, & Trajtenberg, 2001; Posner, 2000).

The ***independent variables*** are *paired patent* and *patenting policy*. We recognize in our data the presence of a *paired patent* to a corresponding scientific paper, indicating that the same piece of scientific knowledge is captured and disclosed in both the form of a scientific paper and a formal patent (which is only possible in a genome center that allows for patenting). This is an important feature known as a patent-paper pair (Ducor, 2000; Murray, 2002; Huang and Murray, in press). We code *paired patent* as a binary variable (1 to denote if a scientific paper is matched to a patent pair and 0 otherwise). To examine the effect of an organizational policy allowing for patenting, we also include *patenting policy* and code it as a binary variable – 1 denotes that a genome center allows for patents (i.e. *Whitehead*, *Baylor* or *Celera*) and 0 otherwise.

We also include a number of ***control variables***: *number of authors*, *article type* (binary variable 1 to denote if the paper is an article or review and 0 otherwise), *journal impact factor* (to account

for the quality of publication),[32] *number of centers* (that collaborate on a publication), and *industry collaboration* (binary variable 1 to denote if the paper results from the collaborative effort of the genome center and a private firm and 0 if it doesn't). In addition, we include dummy variables for each *publication year* (to account for unobserved heterogeneity across each publication cohort) and for each *publication genome center*, namely Whitehead, Sanger, Washington University, Baylor, University of Washington, Stanford, JGI and Celera, (to account for unobserved heterogeneity across each center where the publications come from). Finally, we differentiate and control for the knowledge characteristics of the HGP publications in terms of the type of research output they capture: (i) specific human gene sequence – chromosome and genes including characterization of different types of genes like disease genes or DNA; (ii) gene sequences of other organisms – mouse, rat, zebrafish, worm, and bacteria, including their disease or chromosome-specific characterization; and (iii) techniques, methods or procedures on sequencing or sequencing tools including both hardware and software. To do this, we coded the following binary variables (1 denoting yes and 0 otherwise) for each publication accordingly: (i) *specific human gene sequencing*; (ii) *research on human or application to human*; and (iii) *techniques, methods or tools*.

As verification of genome center patent policy and to match a patent to its corresponding paired scientific paper, we obtain **patent data** for the Whitehead/MIT Center, Baylor and Celera Genomics from the USPTO, crossed checked by patenting and licensing associates from the relevant genome centers. Each patent entry includes patent number; title; inventor names; number of inventors; patent publication date; patent publication country; assignee names; assignee location; assignee code; patent application number; and patent application date.

In addition, we collect the **commercialization data** of the Whitehead Institute based on its HGP patents from the MIT Technology Licensing Office (TLO). The commercialization data includes the number of cases with agreements (including licensing or joint venture); number of cases to start-ups; and start-up company names. As an overview, Whitehead Institute generated about 101

[32] As the impact factor (and the rank ordering) across the set of journals in our data set is stable over time, we used 2003, the last paper publication year in our sample. All genome centers (including Celera) have a significant proportion (about 13% to 30%) of scientific publications in the top 10 journals (i.e. impact factor greater than 25) such as "Science" and "Nature". This speaks to the high quality of research and consistency in publication quality across the centers.

patents in the period 1990 to 2003, of which 40 belong to genome research; only 12 of the 40 patnets are specifically related to the human genome project. They cover a wide range of areas, from gene sequencing techniques, technological tools to genes themselves. Table 6 shows the growing trend of patenting activities from 1990 to 2003, including the HGP-related patents.

Insert Table 6 about here

In addition, based on MIT TLO data, there have been more than 60 licenses to these patents, and the licensees include Alnylam Pharmaceuticals, Ariad Pharmaceuticals, Cell Genesys, Genitrix, LLC, Microbia, Inc., Millennium Pharmaceuticals and Noxxon Pharma AG. Whitehead HGP technologies have also resulted in biotechnology start-ups such as Agencourt Bioscience, which specializes in DNA sequencing and genomic services.

The Whitehead Institute commercialization data covers all the commercialization efforts (beyond patents) from the public consortium as Baylor did not have any start-ups based on their patents from the HGP until the end of 2003. Together with the complete publication and patent data from the genome centers, we have a comprehensive and nuanced data set for our analyses.

4.2. Model specification and estimation

As the dependent variable, *number of cumulative forward citations* to a scientific paper is a highly right-skewed count variable that takes on non-negative integer values. We use a nonlinear regression approach to avoid heteroskedastic, non-normal residuals (Hausman et al., 1984). Furthermore, the dependent variable exhibits over-dispersion with conditional variance significantly greater than the conditional mean (Cameron and Trivedi, 1998).[33] Therefore, we choose negative binomial regression models (NBRM) over Poisson regression models in our estimation, as NBRM overcomes the problem of over-dispersion by assuming a gamma distribution for the conditional mean of the dependent count variable and allows the conditional mean and variance to vary.[34] This choice is consistent with previous works employing citation

[33] This is supported by the likelihood-ratio test where H_1: $E(y_{it})<Var(y_{it})$.

[34] The use of negative binomial regression model is further supported by the results from the goodness-of-fit test which rejected the Poisson distribution assumption.

based measures as dependent variables (Zeidonis, 2004; Hoetker and Agarwal, 2007; Murray and Stern, 2007; Huang and Murray, in press).[35]

Based on this, we developed the following specifications. Equation (1) shows the baseline specification (with controls only) for the negative binomial regression model with robust standard errors estimate:[36]

$$\mathbf{CFC_i = f(\varepsilon_i; \delta NUMBER\_OF\_AUTHORS_i + \nu ARTICLE\_TYPE_i + \mu JOURNAL\_IMPACT\_FACTOR_i + \varphi NUMBER\_OF\_CENTERS_i + \sigma INDUSTRY\_COLLABORATION_i + \chi HGP\_OUTPUT\_CHARACTERISTICS_i)} \quad \mathbf{(1)}$$

From the baseline model, it is possible to develop two further specifications. In addition to the baseline controls specified in Equation (1), the marginal effects model in Equation (2) includes the independent variables *paired patent* and *patenting policy*.

$$\mathbf{CFC_i = f(\varepsilon_i; \alpha PAIRED\_PATENT_i + \beta PATENTING\_POLICY_i + \delta NUMBER\_OF\_AUTHORS_i + \nu ARTICLE\_TYPE_i + \mu JOURNAL\_IMPACT\_FACTOR_i + \varphi NUMBER\_OF\_CENTERS_i + \sigma INDUSTRY\_COLLABORATION_i + \chi HGP\_OUTPUT\_CHARACTERISTICS_i)} \quad \mathbf{(2)}$$

Finally, the full fixed-effects negative binomial regression model in Equation (3) incorporates both publication year fixed effects and center fixed effects to account for unobserved heterogeneities across paper publication cohorts (i.e. each publication year) and across individual genome centers, respectively.

$$\mathbf{CFC_i = f(\varepsilon_i; \alpha PAIRED\_PATENT_i + \beta PATENTING\_POLICY_i + \delta NUMBER\_OF\_AUTHORS_i + \nu ARTICLE\_TYPE_i + \mu JOURNAL\_IMPACT\_FACTOR_i + \varphi NUMBER\_OF\_CENTERS_i + \sigma INDUSTRY\_COLLABORATION_i + \chi HGP\_OUTPUT\_CHARACTERISTICS_i + \eta PUBLICATION\_YEAR\ Fixed\ Effects,t + \psi Center\ Fixed\ Effects,i)} \quad \mathbf{(3)}$$

[35] Alternative Poisson regression models with robust standard errors produced similar results.
[36] We employ the robust standard errors, using Huber-White sandwich estimator (Allison and Waterman, 2002; Greene, 2004) to account for possible heteroscedasticity, and lack of normality in the error terms.

4.3. Empirical results

Our first analysis focuses on how a patenting policy resulting in paired patents relates to the importance of scientific knowledge and genomic innovation. In all our models, we report the coefficients as incidence rate ratios (IRR), which can be derived by exponentiating the coefficients, β_k of the independent variable x_k of the negative binomial regression models. In our case, the IRR can be interpreted as the factor change in annual citations received in a given year due to a unit increase in the regressor. For example, an IRR of 2.03 in the coefficient indicates a 103% increase in the dependent variable for a unit increase in the independent variable, all else being equal.

Table 7 shows the impact of *paired patent* and *patenting policy* on knowledge accumulation. Model 7-1 shows the baseline regression model with controls only. Model 7-2 shows the marginal effects regression model with the independent variables *paired patent* and *patenting policy* included. Model 7-3 shows the full negative binomial regression model (with robust standard errors) incorporating publication year fixed effects and center fixed effects. As a robustness check, the fixed-effects Poisson regression model (with robust standard errors) shown in Model 7-4 produced similar results to Model 7-3.[37]

Insert Table 7 about here

The coefficients and statistical significance of the independent and control variables remain stable and similar across Models 7-1, 7-2 and 7-3.[38] In the most stringent model, Model 7-3, which includes publication year and center fixed effects, we find that publications with a *"paired patent"* is significant (at the 1.6% level), strongly and positively associated with the number of cumulative forward citations. Specifically, a scientific paper with a paired patent is associated

[37] The Poisson model provides a consistent estimate of the conditional mean function, even if the variances are mis-specified (Wooldridge, 1999). However, when there is over-dispersion (as in this case), the Poisson process may result in consistent but inefficient estimates. On the other hand, negative binomial (fixed effects) should yield consistent but efficient estimates in the case of well-specified conditional variance. To assure the readers, fixed-effects Poisson regression model (with robust standard errors) was performed as a robustness check and yielded similar results.

[38] An alternative specification of fixed-effects negative binomial models with robust standard errors, clustered by the eight genome centers to account for possible correlations in the errors for publications within each center produced only slight variation in the magnitude of the standard errors but no change in statistical significance and results across all variables.

with an increase in cumulative citations by more than a factor of two (factor of 2.03). While we cannot be certain that the patent serves as a signal (Hsu and Ziedonis 2007) to enhance the *cumulative* forward citations on a piece of scientific knowledge or innovation over its lifetime, this finding suggests that an organizational pro-patenting policy generates patents that are associated with the most highly cited and important scientific findings and innovations.

Furthermore, publications from centers with a *patenting policy* (i.e. *Whitehead*, *Baylor* and *Celera Genomics*) accrue significantly (at the 1% level) more cumulative citations by a factor of 1.63, relative to the other genome centers. This further supports the notion that an institutional pro-patenting policy enhances the visibility and influence of its scientific research and publications.

In terms of the type of the scientific research, *specific human gene sequencing* variable is significant and negative (by a factor of 0.63), while *techniques, methods or tools* variable is significant and strongly positive (by a factor of 2.28). This suggests that if a publication captures research on specific human gene sequencing information, it is not as highly cited or influential as those potentially more practical and industry driven scientific applications in techniques, methods or tools.

The other control variables largely behaved as expected. The *number of authors* is significant (at the 0.1% level) and positive. This suggests that the higher the number of authors, the higher the cumulative forward citations. While the magnitude of the increase is small (adding another author increases the forward citations by just 1%), the relatively small mean number of authors (about 15), with the standard deviation more than twice as large (about 31), suggest that a more substantial change in the percentage of authors would have a correspondingly larger effect on citations. More authors could signal larger scale or more complex research projects that appeal to a wider network of scientists and hence may be cited more. *Article type* is significant and strongly positive. This is expected because publications that are classified as "articles" or "reviews" are the most noticeable forms of knowledge outputs and therefore most highly cited among other types of publications. *Journal impact factor* is significant and positive. This is expected as it measures a journal's relative importance and "visibility" and provides a proxy for

its quality, especially in comparison with others in the same field. A high impact journal is also one that is more frequently cited. *Number of centers* collaborating on a particular scientific innovation is significant (at the 0.1% level) and positive. This suggests the higher the number of collaborating institutions, the higher the cumulative forward citations. This is reasonable for similar reason as *number of authors*. Finally, *industry collaboration* is not significant. This suggests that teaming up with industry scientists in genomic research and publication may not increase cumulative forward citations to the publications.

4.4. Implications

Through interviews, quantitative analyses and qualitative evidence, we find that patents generated by the genome centers with a patent policy are associated with the most important and influential scientific publications (measured by *cumulative* citations to these publications) at both patent- and organizational- levels of observation. These paired patents lead to agreements, licensing and start-ups. Focusing on organizational governance in IP and disclosure, if the government organizes and funds a variety of scientific projects but allows individual organizations to control their own IP decisions (with little or no regulatory oversight), there are more incentives for instituting a patenting policy from a commercialization perspective, as in the cases of Whitehead/MIT and Celera Genomics.

In other words, while patenting is critical for commercialization to occur in the form of licensing and start-ups, a consequence of organizations with a patenting policy is to generate patents that are, *on the average*, associated with the more important pieces of research. This finding is of concern to policy and decision makers because if patent grant has an adverse *temporal* impact to long-run supply of public (genomic) knowledge (Huang and Murray, in press), it implies that follow-on research and innovation on the more important scientific knowledge are more adversely affected than other knowledge. Hence in the long-run, the supply of the more important public knowledge may be reduced. This would be detrimental to both public research efforts into critical disease areas and private firms that rely on the supply of this public knowledge.

As a result of government experimentation, we are better able to understand variations in organizational policy specifically in IP and disclosure. In addition to the diversity of approach introduced by the government (which is beneficial to the economy), this natural experiment setting has allowed for better policy evaluation. Government has the information and tools to intervene in policy settings through targeted restrictions of patenting on important government-funded projects to maximize rapid and full disclosure of important scientific knowledge. This will make important scientific knowledge widely available for firms, organizations and individuals who depend on such public knowledge to further innovate.

5. Roadmap for designing and analyzing experimentation

Reinterpreted through the lens of economic experiments, the human genome project becomes a setting for a much richer understanding of the ways in which organizational and institutional choices shape scientific productivity. Richly described, these organizational dimensions included the presence or absence of intellectual property, co-located versus distributed work, freedom to publish or not and the imposition of particular forms of copyright and trade secrets over knowledge production. Through careful analysis of this experiment we have demonstrated the power of entrepreneurial experimentation as a way to create diversity in talent, approach and organization and as a starting point for more effective program evaluation. The diversity introduced by the government incorporated different characteristics into the organizations undertaking the same scientific project (in this case, we focused on organizational disclosure and IP choices). Consequently, different types of knowledge and associated commercially oriented outputs were generated at different speed even by the genome centers within the same public consortium. Private Celera genomics had demonstrated a shortened timeline and dramatically different organizational IP strategy in completing the project alongside the public centers.

In its narrow construction, our study is part of a broader effort to take advantage of such experiments to explore the ways in which institutional and organizational arrangements influence the productivity of the scientific community (Murray and Stern, 2007; Murray et al., 2009; Williams, 2009). We contribute to an understanding of these issues by showing that an organizational pro-patenting policy tends to generate patents associated with the most important and influential scientific research. To the extent that such patent grant has an adverse *temporal*

impact on the long-run supply of public (genomic) knowledge (Huang and Murray, in press), follow-on research and innovation on the most important scientific knowledge are more adversely affected than other knowledge. While patents are key to commercialization as in the cases of Whitehead Institute and Celera Genomics, having a policy of no patent and full disclosure allows for rapid dissemination of knowledge and encourages downstream research, development and innovations.

Our study also demonstrates the potentially powerful role of the government serving as an entrepreneur by ensuring a wide range of policy experiments, each of which explores a rich and diverse landscape of possible organizational, technical and institutional configurations. By laying the groundwork for such experimentation, the government is facilitating efforts that lie at the very core of entrepreneurship – diversity and experimentation.

In the arena of science policy the government can promote experimentation in the scientific community by reducing its riskiness and ensuring that the learning is captured across projects – not only along the technical dimension but also along critical organizational and institutional dimensions. As the example of the HGP suggests, the particular dimension of experimentation that is most novel in our discussion of science policy is organizational diversity. Rosenberg (1992) notes that critical dimensions include "size, pattern of ownership, product mix, etc" (p. 193). For experiments in R&D we must consider issues of team size, team distribution, and team disciplinary diversity as well as institutional rules regarding incentives for freedom and control. As Aghion, Dewatripont and Stein (2008) have argued, these are crucial features shaping the incentives for scientists to pursue a diverse set of technical paths. Within the scientific community, such experiments may be particularly feasible given the small-scale of typical operations, the degree to which individual scientists lead their laboratories in an autonomous fashion etc. and the lack of strongly bureaucratic organizations.

The entrepreneurial perspective on the role of science policy in government also allows us to reconsider the debate over the rationale for government spending on R&D in areas where private funding may potentially be available (suggesting that public funding was not overcoming a market failure) – as was the case with the Human Genome Project. By reframing the role of

government as an opportunity to experiment with alternative organizational and institutional arrangements, the focus becomes not on duplicative technical investments but instead on investments in alternative (and potentially more productive) organizational and institutional choices.

Overall, our argument suggests that the government must create an environment in which scientists in the public and private sector, university leaders and corporate leaders have strong incentives to experiment with the way in which they attempt to solve the world's most complex questions. In designing and analyzing economic experimentation, the government can and should first proactively *identify opportunities* to seed a variety of entrepreneurial experiments. While the government need not undertake all such experiments directly, it should provide the specific framework and compass to guide individual organizational policies toward varying their technical and institutional approaches, particularly in their R&D efforts. This should be done in addition to provision of any financial resources.

We also suggest that the government transition from being merely an investor in projects to an entrepreneur, actively *organizing a diversity of economic experiments* on a wide variety of science and R&D projects (among others). The results could lead to more and better options in terms of production, innovative technologies, and institutional arrangements. Science policy could even be designed to purposefully build in competitions or competing groups, employing different technical approaches and organizational policies to achieve similar goals.

Finally, government should carefully *assess and evaluate* the outcomes from each of these (intended) experiments. Mechanisms for assessment, data collection and program evaluation should be tactfully built in at the start of the project. Continuous evaluation and assessment are critical at different stages of the project for the purpose of feedback, gate-keeping and interventions when needed. For example, the direction of the project should be altered (or even stopped) when certain benchmarks at a predetermined stage or time are not met. To enhance efficiency, these mechanisms could be aligned into the incentive structure or organizational levers that facilitate the successful completion of the project. The outcome of the evaluations

should act as inputs, not only for interventions but for initiation of new entrepreneurial experimentations.

Figure 3 outlines the three-stage framework to entrepreneurial experiments that government can undertake to foster a diversity of technical, individual, organizational and institutional approaches to a particular problem and its solution.

Insert Figure 3 about here

If well designed, entrepreneurial experimentation by the government through a variety of policies and practices can not only yield much valued diversity but also more effective program analysis. Simply by changing funding mechanisms and by incorporating new goals into existing monitoring mechanisms, the government could spur a wide range of economic experiments in the scientific community and capture key insights in the process. This would provide a less intrusive means to engage in the science of science management. This agenda holds considerable promise for the development of science of science policy grounded in rich evidence and wide-ranging experimentation.

Acknowledgements

We thank the MIT Technology Licensing Office for generously providing access to its commercialization data. Editor Albert Link and anonymous reviewers gave insightful comments and suggestions which benefited earlier drafts of this paper. The first author gratefully acknowledges the financial support provided by a Merck-MIT fellowship. All errors remain our own.

References

Acemoglu, D., Aghion, P., Bursztyn, L., Hemous, D., 2009. The Environment and Directed Technical Change. MIT Working Paper. http://econ-www.mit.edu/files/4383

Acemoglu, D., 2009. A Note on Diversity and Technological Progress. MIT Working Paper. http://econ-www.mit.edu/files/4373

Aghion, P., Dewatripont, M., Stein, J.C., 2008. Academic freedom, private-sector focus, and the process of innovation. RAND Journal of Economics 39 (3), 617-635.

Arrow, K., 1962. Economic welfare and the allocation of resources for invention. In: The Rate and Direction of Inventive Activity. Princeton University Press, Princeton, NJ.

BBC News, 1999. Human Gene Patents Defended. 27 October. http://news.bbc.co.uk/2/hi/science/nature/487773.stm

Bentley, D.R., 1996. Genomic sequence information should be released immediately and freely in the public domain. Science 274, 533-534.

Bush, V., 1945. Science: The endless frontier. United States Government Printing Office, Washington, D.C.

Business Week, 1990. Cover Story – The Genetic Age. 28 May, No. 3161, pp. 68, McGraw-Hill.

Business Week, 1992. Cover Story – This Genetic Map will Lead to a Pot of Gold. 2 March, No. 3254, Pg 74, McGraw-Hill.

Chicago Sunday Times, 1989. Scientists Calling for U.S. to Spend $3 Billion to Identify Human Genes. 8 January.

Collins, F., Galas, D., 1993. A new five-year plan for the U.S. Human Genome Project. Science 262, 43-46.

Foray, D., 2000. Innovation policy experiment at the regional level. Workshop on the regional level of implementation of innovation and education and training policies, 23-24 November, Brussels, Belgium. ftp://ftp.cordis.europa.eu/pub/improving/docs/ser-foray-paper.pdf

Greenstein, S., 2007. Economic Experiments and Neutrality in Internet Access. Working Paper No. 13158, NBER.

Goldfarb, B., Kirsch, D., Miller, D.A., 2007. Was there too little entry during the dot com era? Journal of Financial Economics 86 (1), 100-144.

Groopman, J., 2001. Annals of Medicine: The Thirty Years' War. The New Yorker. 4 June, pp. 52.

Henderson, R., Cockburn, I., 1994. Measuring competence? Exploring firm effects in pharmaceutical research. Strategic Management Journal 15, 63-84.

Horrobin, D.F., 1986. Glittering prizes for research support, Nature 324, 221-221.

Kalil, T., 2006. Prizes for Technological Innovation. Brookings Institution Working Paper No. 8.

Huang, K.G., Murray, F.E., in press. Does patent strategy shape the long-run supply of public knowledge? Evidence from human genetics. Academy of Management Journal.

Huang, K.G., 2009. Knowledge production in innovative firms under uncertain intellectual property conditions. In: Solomon, G.T. (Ed.), Academy of Management Best Paper Proceedings.

Hunkapiller, T., Kaiser, R.J., Koop, B.F., Hood, L., 1991. Large-scale and automated DNA sequence determination. Science 254 (5028), 59-67.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Jensen, K., Murray, F.E., 2005. Intellectual property landscape of the human genome. Science 310 (5746), 239-240.

Jones, B.F., Wuchty, S., Uzzi, B., 2008. Multi-university research teams: shifting impact, geography, and stratification in science. Science 322 (5905), 1259-1262.

Kaplan S., Murray F.E., in press. Entrepreneurs, institutions and the construction of value in biotechnology. Research in the Sociology of Organizations.

Kolata, G., 2009. Grant System Leads Cancer Researchers to Play It Safe. New York Times, June 29.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.*, 2001. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature 409, 860–921.

Lander, E.S., Weinberg, R.A., 2000. Genomics: journey to the center of biology. Science 287 (5459), 1777-1782.

Lane, J., 2009. Science innovation: assessing the impact of science funding. Science 324 (5932), 1273 – 1275.

Link, A., Scott J. T., 2001. Public/private partnerships: stimulating competition in a dynamic market. International Journal of Industrial Organization 19(5), 763-794.

Murray, F., Stern, S., 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. Journal of Economic Behavior and Organization 63(4), 648-687.

Murray F., Aghion, P., Dewatrapont, M., Kolev, J., Stern., S. 2009. Of Mice and Academics: The Role of Openness in Science. MIT Sloan Working Paper.

Nelson, R., 1959. The simple economics of basic scientific research. Journal of Political Economy 67 (3), 297-306.

Nelson, R., 1961. Uncertainty, learning, and the economics of parallel research and development efforts. Review of Economics and Statistics 43, 351-364.

Reuters Health, 2003. Scientists Complete Human Genome Project. 14 April.

Rosenberg, N., 1992. Economic experiments. Industrial and Corporate Change 1, 181–203

Stephan, P.E., 2008. Science and the university: challenges for future research. CESifo Economic Studies 54 (2), 313-324.

Stern, S., 2005. Economic experiments: the role of entrepreneurship in economic prosperity. In: Understanding Entrepreneurship: A Research and Policy Report, Ewing Marion Kauffman Foundation.

The Scientist, 1999. Hot Papers In Genomics: G. Lennon, C. Auffray, M. Polymeropoulos, M.B. Soares, "The I.M.A.G.E. Consortium: An Integrated Molecular Analysis of Genomes and Their Expression," Genomics, 33:1512, 1996. (Cited in more than 290 papers since publication). 15 February, 13(4), 17. http://www.ornl.gov/sci/techresources/meetings/wccs/hot1_990215.html

The Washington Post, 1987. Congress Begins to Consider the Genome Project. 22 December.

The Washington Post, 1989. The Man behind the Double Helix; Gene-buster James Watson Moves on to Biology's Biggest Challenge – Mapping Heredity. 12 September.

The Washington Post, 1990. Cracking the Body's Code. 5 August.

The Wall Street Journal, 1990. Scientists Find First Clues on How Gene Could Cause Nerve Tumors. 10 August.

National Institute of Standards and Technology, 2009. TIP's Competition in 2009. March. http://www.nist.gov/tip/

U.S. Department of Energy, 1990a. 1988 Memorandum Sets Foundation for Interagency Cooperation. Human Genome Program Human Genome News, May, 2 (1). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v2n1/03memo.shtml

U.S. Department of Energy, 1990b. 5-year Plan Goes to Capitol Hill. Human Genome Program Human Genome News, May, 2 (1). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v2n1/04five.shtml http://www.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/summary.shtml#toc

U.S. Department of Energy, 1993. NIH, DOE Guidelines Encourage Sharing of Data, Resources. Human Genome Program Human Genome News, January, 4 (5). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v4n5/04share.shtml

U.S. Department of Energy, 1994. Genetic Map Goal Met Ahead of Schedule. Human Genome Program Human Genome News, November, 6 (4), 1. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/V6N4/MAPGOALS.shtml

U.S. Department of Energy, 1995a. IMAGE Characterizes cDNA Clones. Human Genome Program Human Genome News, March-April, 6 (6). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v6n6/3image.shtml

U.S. Department of Energy, 1995b. High-Resolution Physical Maps of Chromosomes 16 and 19 Completed. Human Genome Program Human Genome News, January-February, 6 (5), 2. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v6n5/2safchrm.shtml

U.S. Department of Energy, 1995c. Groups Publish Detailed Chromosome 22 Map. Human Genome Program Human Genome News, January-February, 6 (5), 14. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v6n5/14chrom2.shtml

U.S. Department of Energy, 1996a. Detailed Human Physical Map Published by Whitehead-MIT. Human Genome Program Human Genome News, January-March, 7 (5). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n5/05detail.shtml

U.S. Department of Energy, 1996b. DOE, NCHGR Issue Human Subject Guidelines. Human Genome Program Human Genome News, July-September, 8 (1). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v8n1/08humans.shtml

U.S. Department of Energy, 1996c. DOE Merges Genome Center Sequencing Efforts. Human Genome Program Human Genome News, October-December, 8 (2). http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v8n2/01doe.shtml

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.*, 2001. The sequence of the human genome. Science 291 (5507), 1304-1351.

Walsh, J. P., Cho, C., Cohen W. M., 2005. View from the bench: patents and material transfers. Science, 309 (5743), 2002-2003.

Williams, H., 2009. Intellectual Property Rights and Innovation: Evidence from the Human Genome. Working paper, Department of Economics, Harvard University.

Wooldridge, J. M., 1999. Distribution-free estimation of some nonlinear panel data models. Journal of Econometrics. 90 (1), 77-97.

Table 1: Key Attributes of the Seven Public Genome Centers

	Official Founding Year/ Leadership	Affiliation(s)	% Human Genome Sequence Completed	Chromosomes Sequenced	Operating Budget/ Funding/ Employment
Whitehead Institute/ MIT Center for Genome Research	1990 Eric Lander	Massachusetts Institute of Technology (MIT)	~30%	9, 13, 17, 18 and Y	$26 million (NIH grant in 1996), $35 million (NHGRI grant in 1999). $45 to $80 million (annually) >250 people
Wellcome Trust Sanger Institute	1993 John Sulston	Wellcome Trust and the UK Medical Research Council (MRC)	>30%	1, 6, 9, 10, 13, 20, 22, X and (part of) 11	$215 million (Cumulative as of April 2003); $430 million (2001-2006)
Washington University Genome Sequencing Center	1993 Robert Waterson, then Richard K. Wilson	Washington University Medical School	~25%	2, 7, 14, 22 and (part of) X. Coordinator for 2, 4, 7 and Y	$29.7 million grant (from NHGRI & NIH). $6.7 million from 1996 over 3 years
Baylor College of Medicine Human Genome Sequencing Center	1996 Richard Gibbs	Baylor College of Medicine	~10%	3, 12 and (part of) X	$1.3 million from 1996 over 3 years
University of Washington Genome Center	1996 Maynard Olson	University of Washington	~5%	(Part of) 1, 3, 7, 14 and 15	$1 million from 1996 over 3 years
Stanford Human Genome Center	1990 (established in UCSF). 1993 (moved to Stanford). Richard Myers	Stanford University	>11% (with DOE JGI)	4, 5, 16 and 19 (with DOE JGI)	$2.5 million from 1996 over 3 years. ~ 50 faculty, researchers and staff
DOE Joint Genome Institute	1997 (Combining LBNL, LLNL & LANL) Eddy Rubin	U.S. Department of Energy (DOE)	>11% (with Stanford HGC)	5, 16, and 19 (with Stanford HGC)	$60 million (annually), 160 employees

Table 2: Organizational Choices Governing Disclosure and IP of the Genome Centers

	Public	**Private**
Allow Patenting	Whitehead Institute, Baylor	Celera Genomics
No Patenting	Sanger Institute (U.K.), Washington University, University of Washington, Stanford, DOE JGI	

Table 3: Variable Definitions

Name	Definition	Source
Dependent Variables		
Cumulative forward citation (CFC)	Number of cumulative citations made by later papers to the (focal) paper previously published	ISI
Independent Variables		
Paired patent	Binary variable (1/0) denoting if a scientific paper is matched to a patent pair	USPTO
Patenting policy	Binary variable (1/0) denoting if a genome center allows for HGP related patents	USPTO/ Genome centers publication and websites
Control Variables		
Number of authors	Number of authors appearing on the paper	ISI
Article type	Binary variable (1/0) denoting if the paper is an article/review	ISI
Journal Impact Factor	Impact factor (2003) of the journal in which the paper is published	ISI/ Journal Citation Report
Number of centers	Number of unique genome centers with addresses appearing on the paper	ISI
Industry collaboration	Binary variable (1/0) denoting at least one private address on the paper	ISI
Publication year	Year in which the paper is published	ISI
Specific human gene sequencing	Binary variable (1/0) denoting if the paper is on specific human gene sequence or characterization of genes	ISI
Research on human or application to human	Binary variable (1/0) denoting if the paper is on gene sequences of human vs. other organisms such as mouse, rat, zebrafish, worm and bacteria (sequenced as part of the HGP)	ISI
Techniques, methods or tools	Binary variable (1/0) denoting if the paper is on sequencing techniques, methods or tools including both hardware and software	ISI
Whitehead	Binary variables (1/0) denoting if the paper is published by Whitehead Institute/ MIT Center for Genome Research	ISI
Sanger	Binary variables (1/0) denoting if the paper is published by Wellcome Trust Sanger Institute	ISI
Washington University	Binary variables (1/0) denoting if the paper is published by Washington University Genome Sequencing Center	ISI
Baylor	Binary variables (1/0) denoting if the paper is published by Baylor College of Medicine Human Genome Sequencing Center	ISI
University of Washington	Binary variables (1/0) denoting if the paper is published by University of Washington Genome Center	ISI
Stanford	Binary variables (1/0) denoting if the paper is published by Stanford Human Genome Center	ISI
JGI	Binary variables (1/0) denoting if the paper is published by DOE Joint Genome Institute	ISI
Celera	Binary variables (1/0) denoting if the paper is published by Celera Genomics	ISI

Table 4: Summary Statistics of Variables

Dependent Variables					
Variable	**n**	**Mean**	**Std. Dev.**	**Min**	**Max**
Cumulative forward citation	1484	84.59	333.57	0	3984
Independent Variables					
Paired patent	1484	0.01	0.09	0	1
Patenting policy	1484	0.36	0.48	0	1
Control Variables					
Number of authors	1484	14.79	31.23	1	274
Article type	1484	0.75	0.43	0	1
Journal Impact Factor	1484	10.50	9.52	0	34.8
Number of centers	1484	5.05	6.65	1	53
Industry collaboration	1484	0.10	0.30	0	1
Specific human gene sequencing	1484	0.19	0.40	0	1
Research on human or application to human	1484	0.74	0.44	0	1
Techniques, methods or tools	1484	0.12	0.32	0	1
Publication year	1484	1999	3.08	1990	2003
Whitehead	1484	0.13	0.34	0	1
Sanger	1484	0.47	0.50	0	1
Washington University	1484	0.08	0.27	0	1
Baylor	1484	0.18	0.38	0	1
University of Washington	1484	0.04	0.20	0	1
Stanford	1484	0.02	0.15	0	1
JGI	1484	0.03	0.18	0	1
Celera	1484	0.05	0.21	0	1

Table 5: Correlation Matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) Cumulative forward citation	1																			
(2) Paired patent	0.078	1																		
(3) Patenting policy	0.048	0.121	1																	
(4) Number of authors	0.700	-0.012	-0.028	1																
(5) Article type	0.138	0.034	-0.011	0.143	1															
(6) Journal Impact Factor	0.361	0.034	0.045	0.462	0.060	1														
(7) Number of centers	0.625	-0.009	0.052	0.818	0.165	0.410	1													
(8) Industry collaboration	0.284	0.046	0.204	0.337	0.040	0.155	0.339	1												
(9) Specific human gene sequencing	0.118	-0.006	-0.026	0.173	-0.051	0.030	0.106	-0.036	1											
(10) Research on human or application to human	-0.021	0.053	0.185	-0.117	-0.170	-0.065	-0.092	-0.064	0.287	1										
(11) Techniques, methods or tools	-0.036	0.060	0.000	-0.097	0.079	-0.163	-0.131	0.048	-0.165	0.155	1									
(12) Publication year	-0.019	-0.045	-0.301	0.111	0.044	0.024	0.095	0.178	-0.072	-0.149	0.059	1								
(13) Whitehead	0.046	0.119	0.527	-0.005	-0.046	0.105	0.061	-0.043	-0.059	0.057	-0.022	-0.009	1							
(14) Sanger	-0.090	-0.085	-0.700	-0.077	-0.054	-0.093	-0.137	-0.208	0.034	-0.123	-0.010	0.215	-0.369	1						
(15) Washington University	0.027	-0.026	-0.218	0.127	0.092	0.092	0.075	0.013	-0.037	-0.101	0.063	0.036	-0.115	-0.274	1					
(16) Baylor	-0.005	0.018	0.617	-0.065	0.012	-0.058	-0.007	-0.092	0.051	0.172	-0.032	-0.450	-0.181	-0.432	-0.135	1				
(17) University of Washington	0.025	-0.019	-0.155	0.016	0.039	0.017	0.036	-0.011	-0.033	0.014	-0.055	-0.001	-0.082	-0.194	-0.061	-0.095	1			
(18) Stanford	0.037	-0.014	-0.114	0.053	-0.028	0.028	0.045	-0.005	0.095	0.060	-0.014	0.016	-0.060	-0.144	-0.045	-0.071	-0.032	1		
(19) JGI	0.025	-0.016	-0.135	0.038	0.023	-0.043	0.054	0.031	-0.011	-0.069	0.005	0.143	-0.071	-0.170	-0.053	-0.083	-0.037	-0.028	1	
(20) Celera	0.042	0.050	0.303	0.059	0.027	0.037	0.031	0.686	-0.056	0.019	0.092	0.140	-0.089	-0.212	-0.066	-0.104	-0.047	-0.035	-0.041	1

Table 6: Whitehead Patents from 1990 to 2003 (Cumulative Forward Citations as of June 2004)

Pat. Grant Year	90	91	92	93	94	95	96	97	98	99	00	01	02	03
# Patents Belong to Genome Research	1	0	0	0	1	0	1	2	5	8	4	4	7	7
# HGP Related Patents	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**2**	**2**	**4**
Mean CFC	0	0	0	0	0	0	0	25	15	5	1	3	0	0
Mean # Inventors	0	0	0	0	0	0	0	2	1	1	2	3	3	4

Table 7: Impact of Paired Patent and Patenting Policy on Cumulative Knowledge Accumulation

	Main Results: **Negative Binomial Regression Model: *DV* = Cumulative Forward Citations** Coefficients reported as incidence rate ratios, IRR			***Robustness Check:*** **Poisson Regression Model: *DV* = Cumulative Forward Citations** Coefficients reported as incidence rate ratios, IRR
	[7-1] Baseline Model with Controls Only	**[7-2] Marginal Effects Model with Independent Variables**	**[7-3] Full Model with Center and Publication Year Fixed Effects**	**[7-4] Full Model with Center and Publication Year Fixed Effects**
Independent Variables				
Paired patent		**2.86*** (0.86)**	**2.03** (0.60)**	**2.36*** (0.75)**
Patenting policy		**1.67*** (0.22)**	**1.63*** (0.31)**	**1.45** (0.25)**
Control Variables				
Number of authors	**1.01** (0.00)**	**1.01*** (0.00)**	**1.01*** (0.00)**	**1.01*** (0.00)**
Article type	**11.4*** (2.29)**	**11.5*** (2.13)**	**13.8*** (1.90)**	**12.2*** (1.96)**
Journal impact factor	**1.08*** (0.01)**	**1.08*** (0.01)**	**1.08*** (0.00)**	**1.07*** (0.01)**
Number of centers	**1.04*** (0.01)**	**1.04*** (0.01)**	**1.04*** (0.01)**	**1.01* (0.01)**
Industry collaboration	**0.84 (0.12)**	**0.68** (0.10)**	**0.93 (0.15)**	**1.61*** (0.23)**
Specific human gene sequencing	**0.65*** (0.08)**	**0.66*** (0.08)**	**0.63*** (0.07)**	**0.78** (0.09)**
Research on human or application to human	**0.96 (0.17)**	**0.89 (0.14)**	**0.95 (0.09)**	**1.31* (0.18)**
Techniques, methods or tools	**1.98*** (0.39)**	**2.21*** (0.46)**	**2.28*** (0.41)**	**1.95*** (0.38)**
Center fixed effects			Yes	Yes
Publication year fixed effects			Yes	Yes
Regression Statistics				
Log-likelihood	-6198	-6168	-6009	-46371
Wald chi-square (p)	0.000	0.000	0.000	0.000
Number of observations	1484	1484	1484	1484
Robust standard errors (of the IRR) in parentheses. *p<0.10; **p<0.05; ***p<0.01				

Figure 1: Organization of the Human Genome Project: Mapping the Key Stakeholders

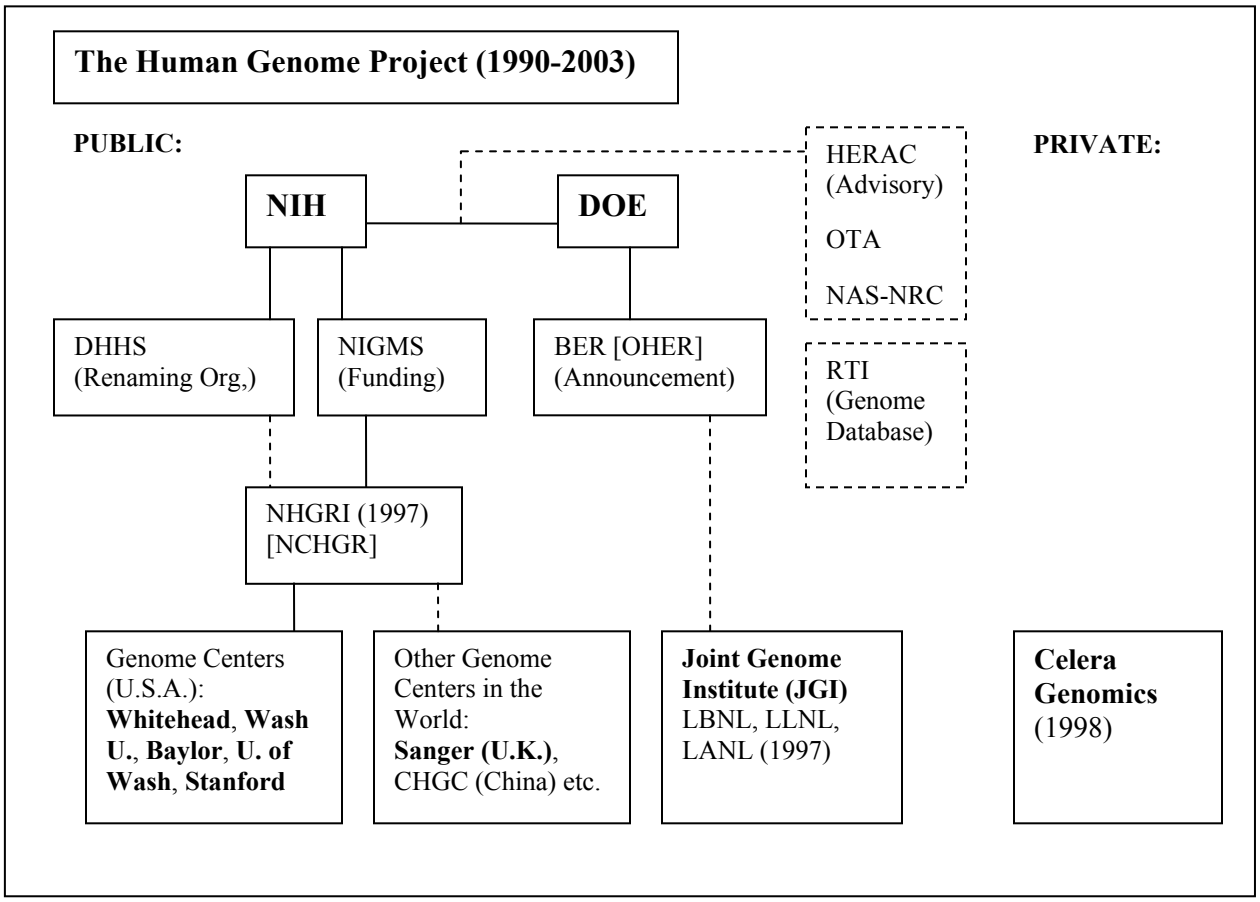

Figure 2: Human Genome Project Publications from Whitehead, Sanger, Washington University, Baylor, University of Washington, Stanford, Joint Genome Institute and Celera Genomics

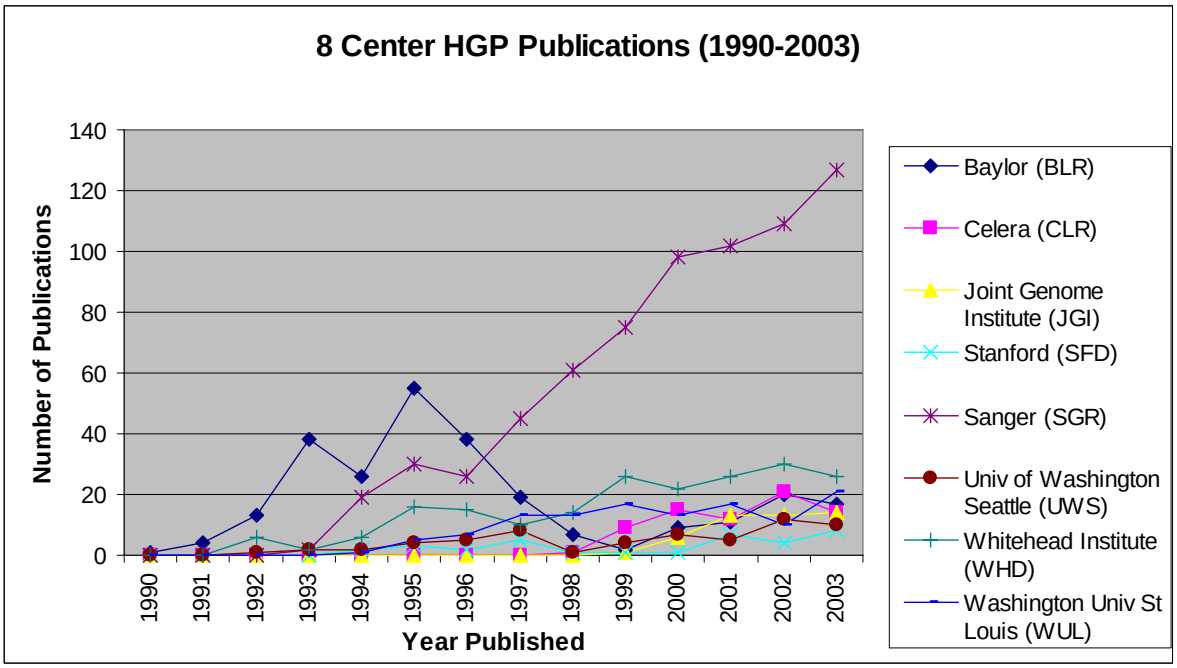

Figure 3: A Three-Stage Framework to Entrepreneurial Experimentations by the Government